Semi-automatic core sentence analysis: improving content analysis for electoral campaign research

Wüest, Bruno ; Clematide, Simon ; Bünzli, Alexandra ; Laupper, Daniel

Abstract: Most automated procedures used for the analysis of textual data do not apply natural language processing techniques. While these applications usually allow for an efficient data collection, most have difficulties to achieve sufficient accuracy because of the high complexity and interdependence of semantic concepts used in the social sciences. Manual content analysis approaches sometimes lack accuracy too, but, more virulently, human coding entails a heavy workload for the researcher. To address this high cost problem without running into the risk of oversimplification, we suggest a semi-automatic approach. Our application implements an innovative coding method based on computational linguistic techniques, i.e. mainly named entity recognition and concept identification. In order to show the potential of this new method, we apply it to an analysis of electoral campaigns. In the first stage of this contribution, we describe how relations between political parties and issues can be recognized by an automated system. In the second stage, we discuss facilities to manually attribute a positive or negative direction to these relations.

Other titles: Electoral campaign and relation mining: extracting semantic network data from swiss newspaper articles

Originally published at:
Wüest, Bruno; Clematide, Simon; Bünzli, Alexandra; Laupper, Daniel (2010). Semi-automatic core sentence analysis: improving content analysis for electoral campaign research. In: JMCE/RECON Workshop 'Computer-aided methods of textual analysis', Berlin, 27 May 2010 - 28 May 2010.

Semi-Automatic Core Sentence Analysis: Improving Content Analysis for Electoral Campaign Research [1]

Bruno Wüest, Simon Clematide, Alexandra Bünzli, Daniel Laupper

Abstract

Most automated procedures used for the analysis of textual data do not apply natural language processing techniques. While these applications usually allow for an efficient data collection, most have difficulties to achieve sufficient accuracy because of the high complexity and interdependence of semantic concepts used in the social sciences. Manual content analysis approaches sometimes lack accuracy too, but, more virulently, human coding entails a heavy workload for the researcher. To address this high cost problem without running into the risk of oversimplification, we suggest a semi-automatic approach. Our application implements an innovative coding method based on computational linguistic techniques, i.e. mainly named entity recognition and concept identification. In order to show the potential of this new method, we apply it to an analysis of electoral campaigns. In the first stage of this contribution, we describe how relations between political parties and issues can be recognized by an automated system. In the second stage, we discuss facilities to manually attribute a positive or negative direction to these relations.

Keywords

Core Sentence Analysis, computer-assisted content analysis, election campaign, natural language processing, relation mining

Word count: 6'377 words (including tables and references)

[1] We thank Hanspeter Kriesi, Michael Hess, Edgar Grande, Lukas Rieder, Maël Mettler, Laura Giess, Dominic Hoeglinger, Marc Helbling, Romain Lachat, Swen Hutter, Martin Dolezal, Stefani Gerber, and the participants of the workshop on „Computer- and Corpus-Linguistic Methods for Large-n Text Analysis in the Social Sciences" at the Freie Universität Berlin for their help and comments.

1. Introduction

Since decades, social scientists have been coding written documents in order to gain empirical insights. While first content analyses were completely conducted manually, automated approaches have recently become much more prominent for data collections. Both manual and automated procedures, however, entail various problems. The most severe drawback of manual coding is the enormous effort of time and costs needed (Schrodt 2009; Hillard et al. 2007). At the same time, the quality of the data gathered is often difficult to assess, mainly as a result of rather low inter- and intra-coder reliabilities or problematic construct validities of theoretically complex variables. In the context of electoral campaigns studies, the Comparative Manifesto Project (CMP) is probably the most renowned example (Budge et al. 2001). Although CMP data is the primary source to measure policy positions of political parties, they suffer from a systematically low reliability (Mikhaylov et al. 2008).[2]

The flaws of traditional coding methods are becoming ever more severe with the emergence of 'big data', i.e. the increasing availability of documents which are interesting for scientific research. The rising public pressure for more transparency of political institutions and the growth of the internet have led to a fast proliferation of digitally available documents (Cortada 2009; Fung et al. 2007). The dominant paradigm in political science to handle 'big data' – especially in the context of party competition – is the estimation of semantic information in documents by means of statistical procedures (Hopkins/King 2007; Laver et al. 2003; Zuell/Landmann 2005; Hillard et al. 2007).[3] A common feature of these approaches is that they code articles using one specific variable, be it the left-right scale, issue categories or ordinal variables. Such procedures either rely on the comparison of relative word frequencies (Laver et al. 2003; Zuell/Landmann 2005; Hillard et al. 2007), the co-occurrence of a few keywords (Ruigrok/van Atteveldt 2007) or on dichotomous variables assessing the presence of word stems (Hopkins/King 2007).

---

[2] A brief explanation of terms which are potentially unfamiliar to the reader can be found in the glossary at the end of this paper.

[3] There is another interesting line of applied computational linguistics in political science. The Kansas Event Data System (KEDS) (Schrodt et al. 1994), its successor TABARI and the further developed version called VRA-Reader (King/Lowe 2003) all apply relation-oriented procedures to identify event data. They process newswire leads and search for sources, actions and targets, which are stored in large dictionaries. The most serious drawback of such software is that it depends on the highly standardized language of press reports, because it is not able to parse natural language. For continuously updated information see Neuendorf/Skalski (2010).

One impressive advantage of these approaches is their independence from costly resources like large keyword dictionaries. All these procedures need is a sufficient amount of manually coded reference texts, usually not more than 100 documents, and they can basically process all kinds of unstructured texts (Hopkins/King 2007: 4f). However, this doesn't hold without restrictions. Changes of vocabulary over time and between different authors as well as large differences in the length of the documents can lead to imprecise coding assessments (Hug/Schulz 2007). The decisive disadvantage for us is that these methods do not meet our primary research goal: the recognition of relational data. On the one hand, some approaches simply do not aim at linking issues to specific actors but try to classify texts (e.g. Hillard et al. 2007). On the other hand, approaches that seek to generate relational data have mainly been used to code party manifestos, parliamentary speeches, or weblogs for which the actors are already pre-defined. Yet, the simultaneous occurrence of multiple actors and issues as well as contradicting political positions of the same actor are not rare in newspapers. For example, an article may cover the electoral campaign efforts of several parties. Another article may instead focus on one party but discusses deviating statements of its exponents on the same policy. Here, a validity problem appears if conventional automated procedures are used.

To gather relational data from newspaper articles, we make use of promising advancements in the field of computational linguistics.[4] Computational linguists have shown that named entity recognition, concept identification and syntactic analysis help to find relations between specific entities in content analyses for social science research (see van Atteveldt et al. 2008). Essentially, we have designed an iterative approach which involes a continuous interaction between the human coder and automated recognition procedures. To put it simply, we exploit computerized schemes to enhance efficiency and reliability and, at the same time, make use of manual coding procedures to increase the validity of our findings. In the following we will outline our conceptual framework, discuss the technical and linguistic implementation and present an exemplary formal evaluation of our approach.

2. An integral measure of party competition

Our starting point was the aim to improve an innovative data collection method for measuring party positions.[5] This approach, the Core Sentence Analysis (CSA), has its origins

[4] Jurafski/Martin (2000); Rinaldi et al. (2005); Porter et al. (2007); West (2001); Evans (2001).

[5] For a comprehensive description of the manual data collection and analysis see Kriesi et al. (2008).

in early theoretical elaborations by Wittgenstein (1984 [1921]) and was first implemented into concrete coding instructions by Osgood (1956) and Axelrod (1976). Recently, it has been renewed for the analysis of electoral campaigns and political conflicts in general (Kriesi et al. 2008; Kleinnijenhuis et al. 1997). Additionally, Franzosi (2004: 60f) has provided theoretical and empirical evidence that the method – he calls it 'story grammars' – is a useful device for the social sciences in general. CSA is an inductive approach which tries to capture the full complexity of a political debate without having to impose theoretical expectations on the data, which constitutes a common problem for content analysis.

The basic idea of this method is that the content of every written document can be described as a network of objects. In our case, we identify relationships between 'political objects', i.e. between a political actor and a political issue (see t*able 1*). Each sentence of a document is reduced to its most basic semantic structure (the so-called core sentence), consisting of a logical subject (*actor, which is either a party or a politician*), its logical object (*issue*) and the direction of the relationship (*polarity*) between the two (using a scale ranging from −1 to +1 with three intermediary positions).

Table 1: Example of a core sentence annotation

Die FDP ist ohne Wenn und Aber für Steuersenkungen. *[The FDP is without ifs and buts for tax reductions]* (Blick, October 4, 2003)		
Subject	**Polarity**	**Object**
FDP	*+ 1*	*tax reduction (budgetary rigor)*

If the parties and politicians as well as the issues are aggregated into meaningful categories, an election campaign can be mapped and evaluated by constructing a network of positions and saliencies. With respect to political actors, every occurrence of a politician or a party is considered relevant, as long as the occurrence is related to an issue position. The aggregation is done by simply summarizing all statements from the same party.[6] More specifically, an actor's position is calculated by taking the average of all statements of one party towards a specific issue, while the saliencies indicate the relative frequency of statements by a party on this specific issue. The issue categories used here were both deductively conceptualized and inductively designed within a large research project, which analyzed the conflict structure of electoral campaigns in Western Europe (see Kriesi et al. 2008 and t*able A.1* in the appendix). As a result, every issue constitutes a consistent

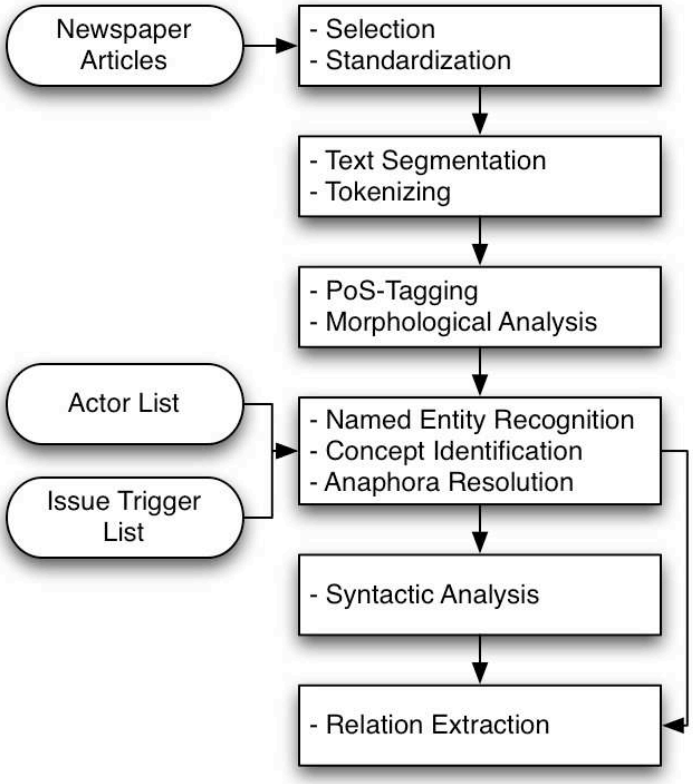
[6] The exact time frame for the selection of articles depends on the amount of material needed to solidly establish the actor positions. Usually, an eight week period is chosen to get enough relevant articles.

aggregate of a conflictive policy field.[7]

3. Implementing the automatic CSA approach

In this section, we present the technical implementation of the new semi-automatic coding method. For the processing and coding of the articles, we used and further developed various linguistic tools. *Figure 1* presents an overview of our text analysis pipeline that integrates the different processing steps by means of a standardized XML format. Accordingly, during the whole processing, an XML file containing the intermediary results of each step can be produced for controls or further analysis.

Figure 1: Software pipeline



Newspaper articles serve as input to the pipeline, the output is the processed article with its metadata (name of newspaper, date, length, title etc.), the annotated entities (political actors and issues) and their recognized relations extracted. The pipeline is designed in a modular way, in order to be able to process different languages and use different tools. Thus,

[7] We, however, use a slightly different categorization than usually employed in CSA studies (see Kriesi et al. 2008): The issues *security* and *army*, *culture* and *cultural liberalism*, and *economic liberalism* and *budgetary rigor* are each subsumed into one category because these categories are difficult to discern in the automated issue detection. All of these category pairs, however, are similar both in terms of their content and with respect to their location in political spaces.

it is possible to adapt it to other languages without too much effort. For this paper, however, only a German version was applied.

3.1. Standardization

At the beginning, the articles are converted and normalized into our pre-defined XML representation, since they originate from different sources like CD-ROMs or digital archives of the publishing houses. Important metadata like the newspaper title, publishing date, length, rubric, and author is encoded in a uniform manner.

3.2. Text segmentation and tokenizing

To split a running text into tokens and sentences, we adapted the tokenizer developed at the University of Stuttgart by Schmid (1994). This step generates a unique ID for every token and sentence within an article.

3.3. Part-of-Speech tagging and morphological analysis

In order to perform concept identification in German, the base forms of the inflected words, i.e. lemmas are needed. For the lemmatization we applied the morphological analysis tool GERTWOL (Koskeniemmi/Haapalainen 1996). To increase precise lemmatization, the words are first tagged by the TnT-Tagger (Brants 2000).

3.4. Named entity recognition (NER) and concept identification

The next step is the recognition of the politicians and issues of interest.[8] We chose an approach that is based on carefully handcrafted lists because we aimed at high accuracy. The respective gazetteer contains 2'710 persons with information about their party affiliation, gender and VIP status, i.e. whether they are famous or high-ranked politicians. In contrast to less prominent persons, VIPs are often referred to only by their last name. There are about 1'990 different last names in our list, and frequent ones such as "Müller" refer to as much as 30 different politicians. The alias resolution, i.e. the assignment of entities to the correct actor, is done at the level of the whole article: First, every occurrence of a single surname or a combination of a first and last name is identified by matching it to the entries in our list. Second, if we found at least one combination of a first and a last name in an article, all other last name mentions are attached to this actor and the other assignments are thrown away.[9] Third, if there is no combination of a first and last name, single surnames known as VIPs

[8] In the following, both political actors and issues are referenced as entities.

[9] So if, for example, a „Stefan Müller" is identified within an article, all mentions of only „Müller" within the same article are considered to refer to „Stefan Müller" as well.

directly resolve to the corresponding person. Any assignments of this occurrence to other actors are again discarded.

Another task in NER is anaphora resolution. After a first mention by his or her name, a politician is often referred to by pronouns or definite noun phrases. State-of-the-art methods for anaphora resolution in German were established by Wunsch (2010) and Klenner/Ailloud (2009). Since their accuracy is still limited, we restricted the anaphora resolution to a very straightforward procedure: If we encounter the personal pronouns “sie” (she) oder “er” (he) in any grammatical case, and, at the same time, a politician of the corresponding gender is found in the previous sentence, the pronoun is resolved to this politician. We treat cases where “sie” refers to parties (in the meaning of “it”), as well as uses of “wir” (we) for references to the collective actor (e.g. “we, the party...”) that occur quite often in interviews, in the same way.

The identification of political issues needs slightly different methods than the recognition of politicians and parties. The issue gazetteer contains a list of manually built trigger patterns for each issue concept. In the simplest case, this is just the base form of a single word, e.g. the compound “Steuersenkung” (tax reduction). However, more often, one word in isolation is usually too general or too ambiguous for reliable concept identification. In these cases, we complemented a keyword with Boolean combinations to connect them with further keywords in the same sentence. This is called a trigger pattern. Including automatically generated orthographic variants, more than 2’100 trigger patterns have been defined. 1’288 of them consist of a single keyword, 791 include two keywords, and 74 are made of three or more keywords.

Further, some trigger words are ambiguous with regard to our issue categories (see *Table A.1* in the appendix). Such ambiguities are resolved on the document level by selecting the issue category with the maximum number of unambiguous hits in the document.

3.5. Syntactic analysis

The next step involves the parsing of a text, i.e. the identification of each sentence's structure. The weighted constraint dependency grammar (CDG) parser (Foth et al. 2004, 2005) is used for a robust syntactic analysis. Although the parser already has a large lexicon, it had to be further adapted to the specific vocabulary of Swiss politics. One feature of the parser is that it produces dependency trees, which display the argument structures more directly than phrase structure trees. Unfortunately the parsing may take several minutes,

especially with long sentences. Therefore, we set a temporal limit to parse our articles in decent time. The downside of this setup is that the parsing of very long sentences is not feasible any more.[10]

In the next section, we present the results of several extraction methods that make use of the computational applications discussed in this section.

4. Determining the validity of our approach: A formal evaluation

We chose the two most recent Swiss national parliamentary election campaigns in 2003 and 2007 to evaluate our automated approach. More specifically, we consider the election coverage in the boulevard newspaper ‚Blick', the largest non-free daily newspaper in Switzerland, up to two and two and a half months prior to the polling day. One the one hand, this decision was motivated by language consideration. We needed a German speaking country to evaluate our language dependent software. Switzerland was then selected because we are most familiar with this country. This facilitates both the development of gazetteers and linguistic rules as well as the interpretation of the results. One the other hand, we have manually annotated data that has been used for actual research at our disposal (e.g. Kriesi et al. 2008; Helbling et al. 2010). This data serves as a gold standard, against which the automatically coded data can be compared.

To improve our method during the evaluation process as well as to ensure a precise error analysis, we split all articles into a development set (187 articles) and a test set (90 articles). The development set was used to train our computational tools and linguistic rules as well as to refine our lists of politicians, parties and issue trigger patterns. The test set was evaluated only once at the end of the development phase. The quality of the test set codings thus serves as an unbiased benchmark of our method since it is applied to previously untreated data. We determine annotation validity on the level of articles, since it is often unclear to which sentence a political statement belongs. Especially in the context of anaphora and long quotations of political actors, it is difficult to pick one sentence as a relation's source (see van Atteveldt et al. 2008: 436). Accordingly, the manual data in the gold standard is often very imprecise regarding the exact source of a relationship.

In the evaluation, we will assess the reliability of the automated methods for the actor, issue, and relation recognition by recall, precision, and F-score (see van Atteveldt et al. 2008;

[10] Different solution to partially remedy this problem exist, e.g. splitting long sentences into subclauses before parsing, or integrating the results of a fast statistical parser into the CDG system (Foth 2007).

Manning/Schütze 2002): The recall indicates how often an entity found by the manual annotation was also found by the automated method. In contrast, the precision indicates how often the automated method is right when it recognizes an entity. The F-Score is the harmonic mean of recall and precision, i.e. it collapses the two indicators to a general measure of fit by giving both indicators the same weight. All of these measures range from 0 to 1, with 1 meaning perfect congruence of the manual and the automated coding. *Table 2* shows these measures for the best performing approach we tried during our research.[11] For the parties and issues, the best method so far is to include only issues and actors that are located within 3 sentences from each other. As regards the relation detection, we applied a simple distance method: every identified issue is combined with its closest party category. Closeness is measured in token distance and limited to the preceding and following sentence. The number of observations (N) shows the sum of entities recognized by these two methods. The number of observations in the gold standards (N_G) is indicated in brackets.[12]

Table 2: Performance of the actor, issue, and relation recognition

	Recall	**Precision**	**F-score**	**N**
development set (N_G = 633)				
parties	0.64	0.49	0.55	824
issues	0.61	0.55	0.58	699
relations	0.52	0.49	0.50	684
test set (N_G = 238)				
parties	0.50	0.64	0.56	187
issues	0.45	0.65	0.53	167
relations	0.38	0.53	0.44	170

Notes: All measures are calculated using data from both election campaigns (2003 and 2007).

The recall for all approaches is obviously lower in the test set than in the development set, where we had the chance to improve the recognition previous to the evaluation. The precision, however and surprisingly, is higher for the test set data. The F-Scores for both the parties and the issues are 0.55 and 0.58 issues in the development set and 0.56 and 0.53, respectively, in the test set. The definition of a text passage that spans over a few sentences as the unit of measurement thus yields acceptable results in terms of accuracy. Further, these results are better than those similar studies for other languages have found (e.g. van Atteveldt et al. 2008). With respect to the relation detection, the simple distance method performs with an F-score of 0.50 for the developments set and 0.44 for the test set. These numbers are still

[11] Besides the approaches used here , we tried various methods to recognize the entities. The evaluation of these different approaches is the subject of another research paper which is in preparation.
[12] Since the data consist of relations, there is always the same number of actors, issues, and relations.

acceptable but rather low, especially for the test set. However, this changes when we move to the level of analysis, which usually is a whole election campaign. Accordingly, we calculated the correlation of the relation detection between the automated and human coding methods for the test set, which equals a Pearson's R score of 0.79. This is not only quite high but, again, also higher than similar research has achieved.

5. From the level of measurement to the level of analysis: An application oriented evaluation

Assessing the quality of our relational data is only meaningful if this data is of actual use for analyses in the social sciences. Therefore, the value of the data at our level of analysis, i.e. election campaigns, has to be considered. In most general terms, salience data provides information on the weight of issues and parties in electoral campaigns and on the importance of specific issues for single parties. Such information is a requirement for various studies of electoral competition: Theories of selective emphasis (Budge/Farlie 1983), issue ownership (Petrocik 2004), and attention shifts (Riker 1986), for instance, all predominantly formulate expectations regarding the salience of specific issues for parties. Similarly, some important approaches in political communication like agenda setting and priming elaborate hypotheses on how the visibility of actors and issues within the media influences public opinion (see McCombs/Shaw 1972; Behr/Iyengar 1985; Kingdon 1995). Our system is able to collect salience data that may be useful for all of these research fields. Thus, *Table 4* and *Figure 2* give a brief illustration of the benefits and shortcomings of the automatically generated data on the four most important parties in the Swiss national electoral campaigns 2003 and 2007 in the light of the above-mentioned research strands.

Table 4: Party salience in the Swiss national election campaigns 2003 and 2007 (automated data): relative frequencies

Party	in %
Social Democratic Party	27.6
Swiss People's Party	25.6
Liberals	22.5
Christian Democratic People's Party	19.6
Total	95.3

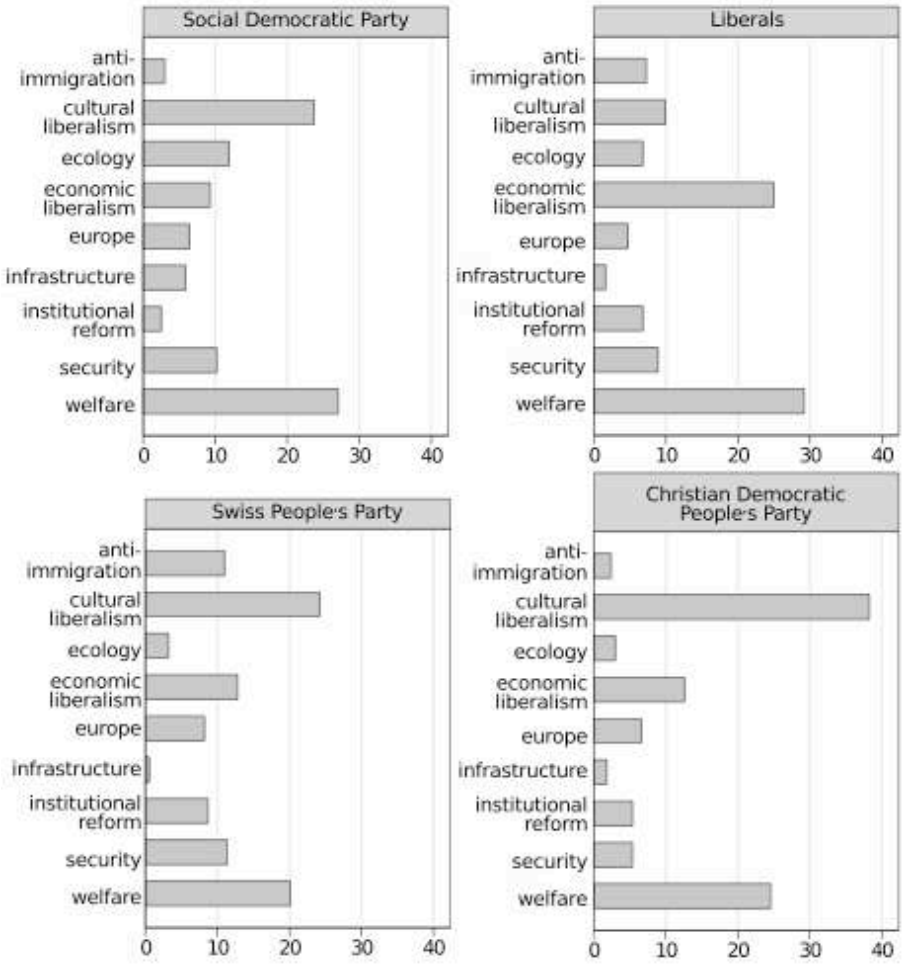
Notes: Only parties with a share of more than 5 percentages in all statements are included into the analysis.

According to the automated content analysis, the Blick's coverage of the election campaign makes the Social Democrats (SPS) the most important party, although it is only the second biggest party with respect to its electoral strength. The biggest party, the populist right-wing Swiss People's Party (SVP), only comes second. This is, on the one hand, a

confirmation of the impression that the Blick is a newspaper with a slight bias to the left (Blum 2005). On the other hand, the result is counterfactual to the often-heard claims that tabloid newspapers strengthen populist parties with their scandalizing and personalizing style of reporting.

Figure 2 shows the parties' relative frequencies of statements about various issues. The Liberals, for instance, try to ‚own' economic issues by emphasizing budgetary rigor and economic liberalism. At the same time, they rarely speak of cultural issues like cultural liberalism. Since the Liberals traditionally see their competence in economic policy making, this is in line with our expectations. The SVP, on the contrary, focuses much more on cultural issues, such as cultural liberalism. Noteworthy is also its emphasis of anti-immigration, an issue all other contestants seem to avoid. Such attention shifts towards ostracized issues has been identified to be an important element of the success of right-wing populist parties in Western Europe (see Kriesi et al. 2008).

Figure 2: Issue salience by party: relative frequencies in %

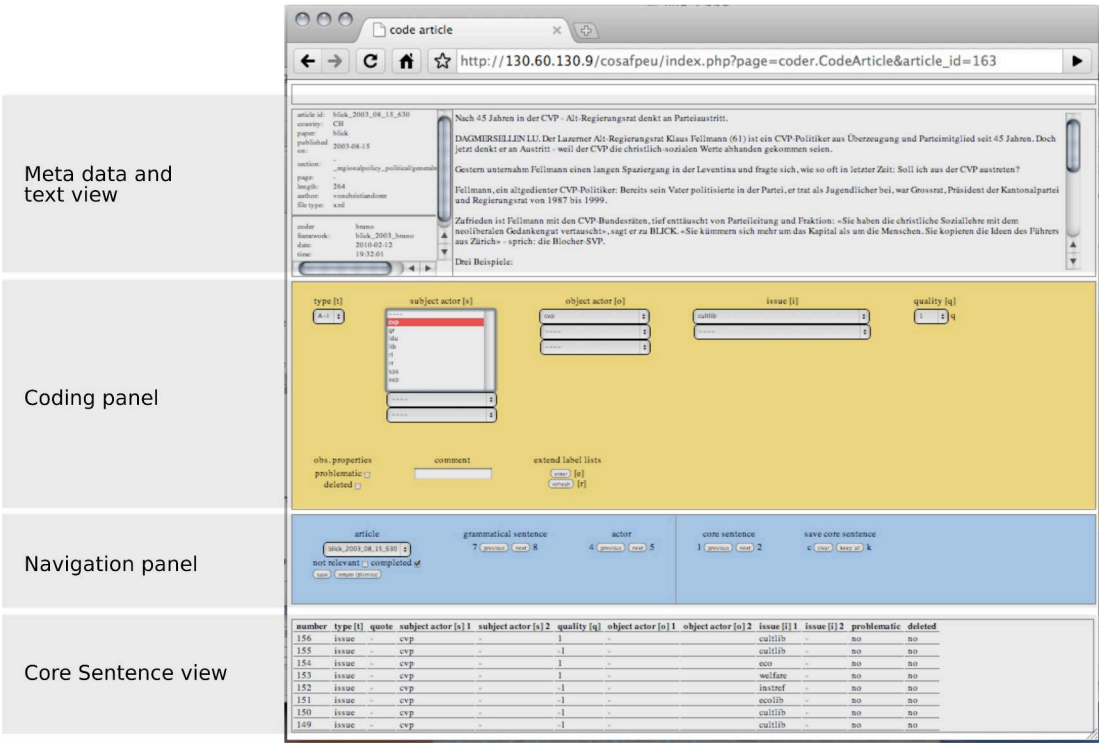


Other issues, like e.g. welfare, are important for all parties. Here, positional data, i.e. information on the different parties' positions with regard to single issues, is crucial to discern their characteristics and strategies. It is of course unlikely that all parties share the same stance on the welfare regime. Thus, if an issue is highly visible and polarized, salience data is only of limited value.

6. Enhancing validity and the recognition of polarity

Polarity is the most important missing dimension in our automated relation detection. Polarity measures capturing whether a relationship between an actor and an issue is positive or negative would provide more detailed insights into the data. In the current stage of our project, we have to add polarity manually. For this task, CoSA, a specifically designed web application for CSA was built.[13] The coding tool is equipped with an administrative panel to organize large-scale data collections and allow simultaneous annotation for up to 10 coders. Further, CoSA includes a database that is specifically adapted to store newspaper articles, metadata on the coding process and the variables. Finally, CoSA comes with an intelligent annotation front end (see *Figure 3*).

Figure 3: Screenshot of the CoSA annotation front end



[13] For further information and download of CoSA see http://www.bruno-wueest.ch/Software.html. All parts of the CoSA framework are open source and, as long as third-party software is not concerned, free to use for scientific purposes. Most programming work on CoSA was done by Stefani Gerber.

The relations found by the automated method can be integrated into the CoSA coding process. They are displayed when the user is navigating to the sentence(s) from where the relation originates. The coders are thus able to control the automated coding and to correct them if necessary. They can also record new relationships the automated system missed. Additionally, human coders are able to add new keywords to the issue and actor gazetteers as they step through the core sentences. Consistent with our interactive approach, coders are thereby providing crucial information for the next relation recognition. Finally, coders can manually determine the polarity of the relations.

Let us briefly return to the example of the welfare issue, which could only be inadequately analyzed in the previous salience analysis. With the help of CoSA, we coded the polarity of each relation we found in the automated relation detection. Since the positions of parties concerning the welfare issue vary considerably, this actually yielded insightful information. To begin with the Social Democrats, they – quite unsurprisingly – stand out by their strong embracement of welfare policies. Their average position is 0.97. The SVP, with an average position of -0.65, on the contrary, fiercely opposes the expansion of the welfare state. The Liberals are slightly objected to welfare, whereas the Christian Democrats (CVP) is clearly in favor of the welfare state. In sum, we have two opposing camps with regard to welfare policies in Switzerland: the centre left camp (CVP and SPS) on the supportive side and the centre right (Liberals and SVP) on the contra side.

7. Concluding remarks

This paper presents a novel approach to (semi-)automatically collect relational data on electoral contests. The Core Sentence Analysis approach can be automated to a certain extent by using computational linguistic tools and techniques. The automated production of data regarding the salience of parties and their issue statements works quite well. Defining a text passage of a few sentences as the unit of measurement offers the best balance between recall and precision at this stage of our research. The remaining inconsistencies of the automated relation recognition can be resolved by having human coders check the results. Additionally, human coders can add polarity to the relations found by the automated approach. The data collected by such an interactive process combines party positions as used in strategic models of party competition in the tradition of Downs (1957) with the concept of salience. Such data is highly demanded by recent literature (see Adams et al. 2005; Meguid 2005).

Our evaluation, however, has shown further need to improve software, linguistic rules and gazetteers to make relation mining a widely accepted approach for content analyses in the social sciences. Furthermore, current methods are adapted to the Swiss context and the processing of the specific vocabulary used by boulevard newspapers. Consequently, future efforts will have to focus on the application of our methods to other newspapers, national settings and languages. Despite the prospect of speeding up the data collection process in comparison with a purely manual approach, human coders are still heavily involved when it comes to the generation of a gazetteer and the recognition of polarity. In exchange, our method produces fine-grained data and is able to do more than text classification. If several hundred newspaper articles of different newspapers are analyzed, this allows us to examine the programmatic supply of parties more precisely than approaches with more common data sets like party manifesto data, expert surveys or roll call data.

References

Adams, J. F./Merrill, S./Grofman, B. 2005: A Unified Theory of Party Competition. A Cross-National Analysis Integrating Spatial and Behavioral Factors, Cambridge.

Axelrod, R. 1976: Structure of Decision. The Cognitive Maps of Political Elites, Princeton.

Behr, R. L./Iyengar, S. 1985: Television news, real-world cues, and changes in the public agenda, in: Public Opinion Quarterly 49, 38–57.

Blum, R. 2005: Politischer Journalismus in der Schweiz, in: Donges, P.(ed.): Politische Kommunikation in der Schweiz, Bern.

Brants, T. 2000: TnT – A Statistical Part-of-Speech Tagger. Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA.

Budge, I./Klingemann, H./Volkens, A./Bara, J./Tanenbaum, E. 2001: Mapping Policy Preferences. Estimates for Parties, Electors, and Governments 1945-1998, Oxford: Oxford University Press.

Budge, I./Farlie, D. 1983: Party Competition – Selective Emphasis or Direct Confrontation? An Alternative View with Data, in: Daalder, H./Mair, P. (eds.): Western European Party Systems. Continuity and Change, London, 267-305.

Cortada, J. 2009: How Societies Embrace Information Technology. Lessons for Management and the Rest of Us, Los Alamitos.

Dolezal, M./Hutter, S./Wueest, B. forthcoming: Exploring the new cleavage across arenas and public debates. Design and methods, in: Kriesi, H./Grande, E./Dolezal, M./Helbling, M./Höglinger, D./Hutter, S./Wüest, B. (eds.): Restructuring political conflict in Western Europe.

Downs, A. 1957: An Economic Theory of Democracy, New York.

Evans, W. 2001: Computer Environments for Content Analysis. Reconceptualizing the Roles of Humans and Computers, in: Burton, O. V. (ed.): Computing in the Social Sciences and Humanities, Urbana, IL, 67-86.

Foth, K./Daum, M./Menzel, W. 2004: A broad coverage parser for German based on defeasible constraints, in: Christiansen, H./Skadhauge, P. R./Villadsen, J. (eds.): Proceedings Constraint Solving and Language Processing. Workshop Proceedings. Datalogiske Skrifter No. 99, Roskilde, 88-101.

Foth, K./Menzel, W./Schröder, I. 2005: Robust parsing with weighted constraints, in: Natural Language Engineering 11/1, 1-25.

Foth, K. A. 2007: Hybrid Methods of Natural Language Analysis, Aachen.

Franzosi, R. 2004: From Words to Numbers. Narrative Data and Social Science, Cambridge.

Fung, A./Graham, M./Weil, D. 2007: Full Disclosure. The Perils and Promise of Transparency, Cambridge.

Helbling, M./Hoeglinger, D./Wueest B. 2010: How Political Parties Frame European Integration, in: European Journal of Political Research 49.

Hillard, D./Purpura, S./Wilkerson J. 2007: An Active Learning Framework for Classifying Political Text. Paper presented at the 2007 Annual Meeting of the Midwest Political Science Association, Chicago.

Hopkins, D./King, G. 2007: Extracting Systematic Social Science Meaning from Text, in: http://people.iq.harvard.edu/~dhopkins/papers.html; Oct. 1, 2007.

Hug, S./Schulz, T. (2007). Left-right Positions of Political Parties in Switzerland, in: Party Politics 13/3, 305-330.

Jurafsky, D./Martin, J. H. 2000: Speech and Language Processing. An Introduction to Natural Language Processing, Computional Linguistics and Speech Recognition, Upper Saddle River, NJ.

Kim, S.-M./Hovy, E. 2004: Determining the sentiment of opinions, in: Proceedings of Coling 2004, Geneva, 1367-1373.

King, G./Lowe, W. 2003: An Automated Tool for International Conflict Data with Performance as Good as Human Coders. A Rare Events Evaluation Design, in: International Organization 57, 617-642.

Kingdon, J. W. 1995: Agendas, alternatives and public policies, Boston.

Kleinnijenhuis, J./de Ridder, J. A./Rietberg, E. M. 1997: Reasoning in Economic Discourse. An Application of the Network Approach to the Dutch Pressin, in: Roberts, C. W. (ed.): Text

analysis for the Social Sciences. Methods for Drawing Statstical Inferences from Texts and Transcripts, Mahwah, NJ, 191-209.

Klenner, M./Fahrni, A./Petrakis, S. 2009: PolArt. A Robust Tool for Sentiment Analysis, in: Jokinen, K./Bick, E. (eds.): Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009, 235-238.

Klenner, M./Ailloud, É. 2009: Optimization in coreference resolution is not needed. A nearly-optimal zero-one ILP algorithm with intensional constraints. 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09).

Koskeniemmi, K./Haapalainen, M. 1996: GERTWOL – Lingsoft Oy, in: Hausser, R. (ed.): Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994, Tübingen, 121-140.

Kriesi, H./Grande E./Lachat, R./Dolezal M./Bornschier S./Frey T. 2008: West European Politics in an Age of Globalization, Cambridge.

Krippendorff, K. 2004: Content Analysis. An Introduction to Its Methodology. 2nd ed, Thousand Oaks.

Laver, M./Benoit, K./Garry, J. 2003: Extracting Policy Positions from Political Texts Using Words as Data, in: American Political Science Review 97/2, 311-331.

Manning, C./Schütze, H. 2002: Foundations of statistical natural language processing, Cambridge.

McCombs, M. E./Shaw, D. L. 1972: The agenda-setting function of the mass media, in: Public Opinion Quarterly 36, 176-187.

Meguid, B. M. 2005: Competition between Unequals. The role of mainstream party strategy in niche party success, in: American Political Science Review 99/3, 347-359.

Mikhaylov, S/ Laver M./Benoit K. 2008: Coder Reliability and Misclassification in Comparative Manifesto Project Codings. Paper prepared for presentation at the 66th MPSA Annual National Conference, Palmer House Hilton Hotel and Towers, April 3–6, 2008.

Neuendorf, K. A./Skalski, P. 2010: The Content Analysis Guidebook Online, in: http://academic.csuohio.edu/kneuendorf/content/; April 9, 2010.

Osgood, C. E. 1956: Evaluative Assertion Analysis, in: Litera 3, 47-102.

Petrocik, J. R. 2004: Issue Ownership and Presidential Campaigning. 1952-2000, in: Political Science Quarterly 118/4, 599-626.

Porter, B./Barker, K./Hovy, E. H. 2007: Learning by Reading. A Prototype System, Performance Baseline and Lessons Learned. Paper presented at the meeting of the Association for the Advancement of Artificial Intelligence, Vancouver.

Riker, W. H. 1986: The Art of Political Manipulation, New Haven.

Rinaldi, F./Schneider G./Kaljurand K./Hess M./Andronis C./Persidis A./Konstanti O. 2005: Relation Mining over a Corpus of Scientific Literature, in: Lecture Notes in Computer Science 3581, 535-544.

Ruigrok, N./van Atteveldt, W. 2007: Global Angling with a Local Angle. How U.S., British, and Dutch Newspapers Frame Global and Local Terrorist Attacks, in: The Harvard International Journal of Press/Politics 12, 68-90.

Schmid, H. 1994: Probabilistic Part-of-Speech Tagging Using Decision Trees. Paper presented at the International Conference on New Methods in Language Processing, Manchester.

Schrodt, P. A. 2009: Reflections on the State of Political Methodology, in: Newsletter of the Political Methodology Section. American Political Science Association 17/1, 1-4.

Schrodt, P. A./Davis, S. G./Weddle, J. L. 1994: Political Science. KEDS – A Program for the Machine Coding of Event Data, in: Social Science Computer Review 38/4, 561-88.

Sennrich, R./Schneider, G./Volk M./Warin M. 2009: A New Hybrid Dependency Parser. Proceedings of GSCL-Conference, Potsdam.

van Atteveldt, W./Kleinnijenhuis, J./Ruigrok, N. 2008: Parsing, Semantic Networks, and Political Authority. Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles, in: Political Analysis 16, 428-446.

West, M. D. 2001: Applications of Computer Content Analysis, Westport.

Wittgenstein, L. 1984 [1921]: Tractatus Logico-Philosophicus, Frankfurt a. M.

Wunsch, H. 2010: Rule-based and Memory-based Pronoun Resolution for German. A Comparison and Assessment of Data Sources. PhD thesis, Universität Tübingen.

Zuell, C./Landmann, J. 2005: Identifying National and International Events Using Computer-Assisted Content Analysis. Paper presented at the first EASR conference in Barcelona, Barcelona.

Glossary

Anaphora are linguistic elements, mostly pronouns or definite nominal phrases, which are referring to other elements in a text.

Gazetteer is a list of names for specific entities held in computer form, which allows for rapid search and query.

Gold standard (in content analysis evaluations) denotes data that was generated by other methods than the evaluated procedure. This data is treated as paragon of excellence against which the new data is compared.

Lemma. A Lemma is the base form of an inflected word form, e.g. 'go' is the lemma of 'went', 'goes', or 'gone'.

Named entity recognition detects proper names, e.g. names of parties or politicians, in text documents. Assigning different name mentions in a text to the same reference entity is called alias resolution.

Parsers compute the syntactical structure of sentences, i.e. which tokens form the subject, verb, objects and so forth of the sentence.

Part-of-Speech is the lexical category of a word. The most important lexical categories are nouns, verbs, adjectives and prepositions.

Reliability is concerned with the question whether the data collection is more stable over time, better reproducible by different coders, and more accurate compared to some canonical standard (see Krippendorff 2004).

Syntactic dependency indicates the syntactic relations between phrases (subject, object, predicate, etc.) in a sentence.

Tagger A tagger marks every token (words and punctuation marks) with a part-of-speech label, a so-called *tag*.

Tokens (in computational linguistics) are a sequence of characters that serve as basic elements for further linguistic processing. Typical tokens are word forms and punctuation marks.

Trigger words are words in the documents that initiate a computational linguistic procedure, e.g. the recognition of concepts.

Validity refers to the question whether the collected data actually measures the theoretically derived concepts.

XML (Extensible Markup Language) is a standard format for encoding documents as structured textual data in machine-readable form.

About the Authors

Bruno Wüest, MA., Visiting scholar at New York University and Researcher at the Center for Comparative and International Studies, University of Zurich; wueest@ipz.uzh.ch. He is interested in industrial relations, economic globalization, party politics and automated content analyses.

Dr. Simon Clematide, Senior researcher at the Institute of Computional Linguistics, University of Zurich; simon.clematide@cl.uzh.ch. He is interested in automatic text analysis, German morphology and syntax, named entity recognition, and interdisciplinary applications of language technology.

Alexandra Bünzli, lic. phil., Researcher at the Institute of Computional Linguistics, University of Zurich; buenzli@cl.uzh.ch. She is interested in automated text analysis, syntax-semantic interfaces, semantic representations, and interdisciplinary natural language processing, particularly with regard to legal texts.

Daniel Laupper, Student in Political Science at the Center for Comparative and International Studies, University of Zurich; daniel.laupper@access.uzh.ch. He is interested in comparative politics, political methodology and automated content analysis.

Appendix

Table A.1: *Categorization of issues (see Dolezal et al. forthcoming)*

Categories	**Description**
Economic liberalism	Opposition to market regulation; opposition to economic protectionism in agriculture and other sectors of the economy.; support for deregulation, more competition, and privatization; support for a rigid budgetary policy; reduction of the state deficit; cuts on expenditures; reduction of taxes without direct effects on redistribution
Anti-immigration	Support for a tough immigration and integration policy
Europe	Support for European integration
Welfare	Support for an expansion of the welfare state; defense against welfare state retrenchment; support for tax reforms with a redistributive character; calls for employment and health care programs
Cultural liberalism	Support for cultural diversity, international cooperation (excluding the European Union and Nato); support for the United Nations; support for the right to abortion and euthanasia; opposition to patriotism, calls for national solidarity, defense of tradition, national sovereignty, and to traditional moral values; support for a liberal drug policy; support for education, culture, and scientific research.
Security	Support for more law-and-order, the fight against crime, and denouncing political corruption; support for the armed forces (including Nato), for a strong national defense, and for nuclear weapons
Ecology	Opposition to nuclear energy; support for environmental protection
Institutional reform	Support for various institutional reforms, i.e. modifications in the structure of the political system
Infrastructure	Support for the improvement of the country's roads, railways, etc.